FugakuNEXT: AI-HPC platform

Advanced Technology Development Unit

Fujitsu Research

January 2026

Fujitsu Challenges on Supercomputer

FUJITSU

Continue to provide cutting-edge supercomputers

The world's fastest
(June 2011: TOP500)
K computer

The world's fastest
(November 2019: Green500,
June 2020: TOP500/HPCG/HPL-AI/Graph500)
Supercomputer Fugaku

FugakuNEXT

PRIMEHPC FX100

PRIMEHPC FX10

PRIMEHPC FX1000/FX700
Using the same A64FX CPU as Fugaku

The world's fastest vector processor
(1999)

VPP5000

SPARC Enterprise

FX1

The world's highest execution efficiency
(November 2008: TOP500)

The world's fastest
(November 1993, TOP500)

Gordon Bell Award
(1994, 95, 96)

NWT

VPP300/700

VPP500

PRIMEPOWER HPC2500

PRIMERGY CX400/600
Cluster node

PRIMERGY BX900
Cluster node

HX600
Cluster node

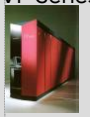
PRIMERGY RX200
Cluster node

Japan's most powerful x86 cluster
(June 2004, TOP500)

The world's highest scalability
(2003)

Vector

VP Series



F230-75APU

Japan's first supercomputers
(1977)

AP3000

AP1000

Scalar

Cluster

TC Cloud
Analytical simulation cloud services (2011)

Cloud

Computing as a Service (CaaS)
Cloud service of Fujitsu's advanced computing technologies including A64FX

~1980 1985 1990 1995 2000 2005 2010 2015 2020 2030

Future Supercomputing

FUJITSU

Trends in HPC Technologies

Ever-Growing data volumes and computation power

Higher computational density and power efficiency

Rapid AI evolution and expansion of applicable areas

AI-HPC platform

- FUJITSU's Arm-based CPU proven in HPC and GPUs optimized for large-scale AI training
- Driving innovation across science, technology, and industry to shape humanity's future

The impact of AI-HPC platform

AI-enhanced Simulation

- Significantly improve accuracy and speed by optimize simulation parameters using AI

Data-Driven Science

- Uncover hidden correlations, causalities, and new scientific laws by analyzing massive datasets

Research Process Optimization

- Accelerate scientific discovery by optimizing research workflows from experiment design to insight

Knowledge Discovery

- Discover new hypotheses and knowledge by mining vast scientific papers and literatures.

FugakuNEXT
– Japan's Flagship Platform driving "AI for Science"

FUJITSU

FugakuNEXT development Goals

Made with Japan	Technological Breakthroughs	Sustainability and Continuity
• Joint development of RIKEN, NVIDIA, and Fujitsu • Create global innovation and collaboration	• Tight integration of CPU–GPU for high-bandwidth, heterogeneous nodes • Achieve over 100× application performance by AI-HPC integration	• A sustainable software ecosystem • Modernization of applications • Advanced energy-efficient operation technology

The FugakuNEXT Ecosystem

- Accelerating scientific progress through AI for Science methodologies
- R&D leadership in advanced computing and AI technologies
- Sustained high-end processor development to secure computational resources

RIKEN
- System Design, Software and Application development

Fujitsu
- CPU and System Development

NVIDIA
- GPU development

Schedule

CY 2025	2026	2027	2028	2029	2030	2031

Basic Design
- Define overall architecture

Detailed Design
- Develop production readiness plan
- Finalize Software/Hardware Specifications

Manufacturing / Installation
- Produce individual parts and modules
- Prepare for user Acceptance Testing

Operation
- Conduct performance improvements and maintenance activities

Reference: MEXT (2025). AI for Science: Concepts and Directions. https://www.mext.go.jp/content/20251006-mxt_jyohoka01-000045188_03.pdf

The world's top-level AI-HPC platform with Fujitsu CPU and NVIDIA GPU

Designed to deliver up to 100× improvement in application performance.

FUJITSU

Specification Comparison

Specification	FugakuNEXT			Fugaku
	CPU	GPU		CPU
Number of Nodes	≧3,400			158,976
FP64 Vector perf.	≧48 PFLOPS	≧2.6 EFLOPS	×4.9	537 PFLOPS
FP16/BF16 Matrix perf.	≧1.5 EFLOPS	≧150 EFLOPS	×70.5	2.15 EFLOPS
FP8 Matrix perf.	≧3.0 EFLOPS	≧300 EFLOPS		-
FP8 Matrix sparsity perf.	-	≧ 600 EFLOPS		-
Memory Capacity	≧10 PiB	≧10 PiB	×4.1	4.85 PiB
Memory BW	≧7 PB/s	≧800 PB/s	×4.9	163 PB/s

FUJITSU-MONAKA-X Processor

FUJITSU

Arm-based Processor Enhanced for AI

FUJITSU-
MONAKA-X

Optimization for HPC

- Next-generation **3D many-core architecture**
- **1.4 nm process** technology
- Further acceleration through **SIMD extensions**
- Legacy support for existing HPC applications

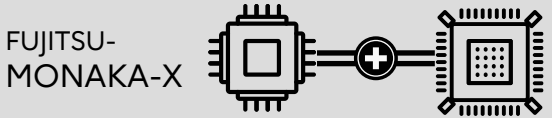
AI Processing Acceleration

- **Arm SME**: Implementation of NPU
- World's first in server-class CPUs
- Enhanced application coverage via **GPU collaboration**

Power efficiency and reliability

- **Ultra-low voltage** operation control
- Enhanced security with **Confidential Computing**
- Robust reliability through **RAS functionality**

CPU-GPU Tight Integration



NVLink™ Fusion

Accelerating AI-HPC Converged Workloads via High-Bandwidth, Low-Latency, and Coherent CPU-GPU Access

CPU Strengths

- complex control flow, latency-sensitive tasks, and irregular memory access

✓ Simulation, Realtime AI, Multimodal processing

+

GPU Strengths

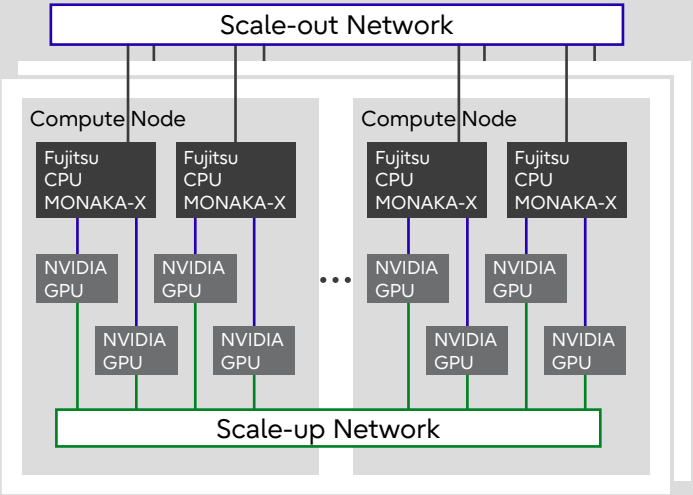
- Massively parallel data-parallel execution with regular memory access patterns

✓ Large scale DL/ML training, Image/signal recognition

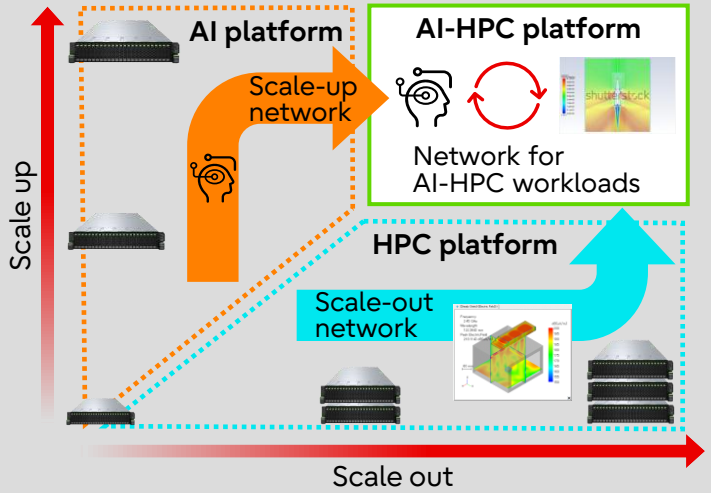
Integrating Scale-up/out network

- Heterogeneous interconnect with Scale-up and Scale-out
- Leverages both scale-up and scale-out capabilities by optimizing at the system level
  - Enhancing AI-HPC integrated workload performance through optimal allocation of network resources
  - Significant performance gains of HPC workloads with effective scale-up network utilization

Heterogeneous Interconnect



Interconnect for AI-HPC platform

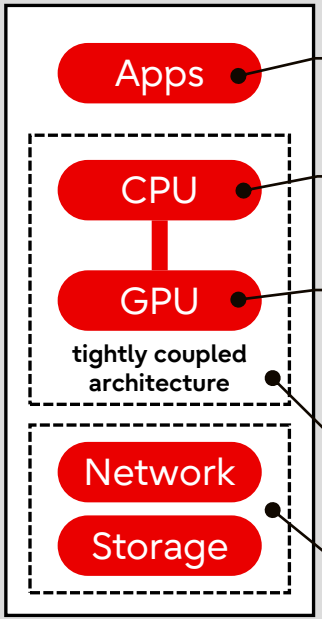


Software-Driven 100x Application Performance Over Fugaku

FUJITSU

- Achieving 100x application performance necessitates software leveraging hardware far more efficiently.
- We believe total performance optimization, from application to system software, can achieve the goal.

Keys for Achieving 100x Application Performance



Leveraging AI to accelerate simulations
- Hybrid surrogate model & simulation, accuracy-preserving quantization

Fully leveraging MONAKA-X of ultra many core for boosting existing HPC applications

Accelerating high-precision operations with low-precision arithmetic, utilizing like Ozaki Scheme

Harnessing CPU-GPU tightly coupled architecture for application speedup
- Optimal workload & data placement on CPU / GPU to maximize their capabilities (e.g. Large node memory, massively parallel SIMD)

Optimizing data communication for resolving data bottleneck

Thank you